Journal of Multivariate Analysis 112 (2012) 207–218

Contents lists available at SciVerse ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Dimension reduction for the conditional kth moment via central solution space

Yuexiao Dong [a,*], Zhou Yu [b]

[a] Department of Statistics, Temple University, Philadelphia, PA, 19122, USA

[b] School of Finance and Statistics, East China Normal University, Shanghai, 200241, China

ARTICLE INFO

Article history:
Received 2 March 2010
Available online 1 July 2012

AMS subject classification:
62B05

Keywords:
Central kth moment space
Central solution space
Dimension reduction space
Non-elliptical distribution

ABSTRACT

Sufficient dimension reduction aims at finding transformations of predictor X without losing any regression information of Y versus X. If we are only interested in the information contained in the mean function or the kth moment function of Y given X, estimation of the central mean space or the central kth moment space becomes our focus. However, existing estimators for the central mean space and the central kth moment space require a linearity assumption on the predictor distribution. In this paper, we relax this stringent assumption via the notion of central kth moment solution space. Simulation studies and analysis of the Massachusetts college data set confirm that our proposed estimators of the central kth moment space outperform existing methods for non-elliptically distributed predictors.

1. Introduction

The scale and complexity of data sets have increased drastically due to the development in modern technology. High-dimensional data pose many challenges for statisticians. Direct application of many nonparametric or semiparametric methods may fail because of the curse of dimensionality. Sufficient dimension reduction [4,7,12,13] aims at preserving all the relevant information for either regression or classification and effectively transforms high dimensional problems to lower dimensions. Let X be a p-dimensional random vector representing the predictor, and Y be a random variable representing the response. Sufficient dimension reduction assumes $Y \perp\!\!\!\perp X|\beta^T X$, where $\perp\!\!\!\perp$ means independence and $\beta \in \mathbb{R}^{p\times d}$ with $d < p$. The column space of β, denoted by Span(β), is a dimension reduction space (DRS). When the intersection of all DRSs is still a DRS, we call this intersection the central space (CS). Discussions about the existence of the central space can be found in [4], and [24] provided a more general result. Denote $\mathcal{S}_{Y|X}$ [1,2] as the central space of Y versus X. The dimension of $\mathcal{S}_{Y|X}$ is the structural dimension.

In many situations, regression analysis is mostly concerned with inferring the conditional mean $E(Y|X)$. Cook and Li [5] introduced the notion of central mean space (CMS), which is denoted by $\mathcal{S}_{E(Y|X)}$. The central mean space is the intersection of Span(β) over all β satisfying $Y \perp\!\!\!\perp E(Y|X)|\beta^T X$. Following the idea of central mean space, central kth moment dimension reduction space (CKMS) [23] focuses on reducing the complexity of the mean function, the variance function, and up to the kth moment function, leaving the rest of the conditional distribution of Y given X as a nuisance parameter. Let $M^{(k)}(Y|X) = E\{[Y - E(Y|X)]^k|X\}$ for $k \geq 2$ and $M^{(1)}(Y|X) = E(Y|X)$. Yin and Cook [23] made the following definition.

Definition 1. If $Y \perp\!\!\!\perp \{M^{(1)}(Y|X), \ldots, M^{(k)}(Y|X)\}|\beta^T X$, then Span($\beta$) is a kth moment DRS for the regression of Y versus X. Let $\mathcal{S}^{(k)}_{Y|X}$ be the intersection of all kth moment DRSs. If $\mathcal{S}^{(k)}_{Y|X}$ is itself a kth moment DRS, then it is called the central kth moment DRS, or CKMS for short.

* Corresponding author.
E-mail address: ydong@temple.edu (Y. Dong).

0047-259X/$ – see front matter
doi:10.1016/j.jmva.2012.06.001

Table 1
Z scale moment-based estimators of CMS, CKMS, and CS.

DRS	First-order methods	Second-order methods
$\mathcal{S}_{E(Y\|Z)}$	$\beta_{\mathrm{OLS}} = E(ZY)$	$\beta_{\mathrm{PHD}} = \Sigma_{yzz}\Sigma_{yzz}^T$, where $\Sigma_{yzz} = E\{[Y - E(Y)]ZZ^T\}$
$\mathcal{S}_{Y\|Z}^{(k)}$	$\beta_{\mathrm{COV}}^{(k)} = E(T_k \otimes Z)$, where $T_k = (Y, Y^2, \ldots, Y^k)$	$\beta_{\mathrm{PHD}}^{(k)} = E\{[T_k - E(T_k)] \otimes (ZZ^T)\}$, where $T_k = (Y, Y^2, \ldots, Y^k)$
$\mathcal{S}_{Y\|Z}$	$\beta_{\mathrm{SIR}} = \mathrm{Var}[E(Z\|Y)]$	$\beta_{\mathrm{SAVE}} = E[I_p - \mathrm{Var}(Z\|Y)]^2$

By definition, CMS is a special case of CKMS with $k = 1$. Because $Y \perp\!\!\!\perp X|\beta^T X$ implies $Y \perp\!\!\!\perp E(Y|X)|\beta^T X$, a DRS is also a mean dimension reduction space. It follows that $\mathcal{S}_{E(Y|X)} \subseteq \mathcal{S}_{Y|X}$, since $\mathcal{S}_{E(Y|X)}$ is the intersection of at least the same, if not more, subspaces. The CKMS is contained in the central space for the same reasoning. We can also see that a kth moment DRS must be an ith moment DRS for any $i \leq k$. These relationships are summarized by Yin and Cook [23] as $\mathcal{S}_{E(Y|X)} \subseteq \cdots \subseteq \mathcal{S}_{Y|X}^{(k)} \subseteq \cdots \subseteq \mathcal{S}_{Y|X}$.

Central mean space and central kth moment space provide more insight into existing dimension reduction methods, such as sliced inverse regression (SIR) [12], sliced average variance estimator (SAVE) [7], ordinary least squares (OLS) [16] and principal Hessian directions (PHD) [3,13]. It is now well known that SIR and SAVE are estimators of the central space, while OLS and PHD are estimators of the central mean space [5]. Extensions of OLS and PHD estimators are developed to estimate the central kth moment space in [23,22] respectively. All the aforementioned methods are based on moments and/or inverse conditional moments. As discussed in [9], estimators that involve linear functions of X, such as $E(XY^k)$, $E(X|Y)$, will be called the first-order methods. Those that involve both linear and quadratic functions of X, such as $E(XY^k)$, $E(X|Y)$, $E(Y^kXX^T)$, $E(XX^T|Y)$, will be referred to as the second-order methods.

Denote the covariance matrix of X as Σ, which is assumed to be nonsingular. The standardized predictor is $Z = \Sigma^{-1/2}[X - E(X)]$. It follows from the definition of the CKMS that $\mathcal{S}_{Y|X}^{(k)} = \Sigma^{-1/2}\mathcal{S}_{Y|Z}^{(k)}$. Similar relationships exist for the CMS's and CS's. Without loss of generality, we will first work with the standardized predictor Z and then transform the result back to the original X scale unless stated otherwise. A non-exhaustive list of moment-based sufficient dimension reduction estimators is summarized in Table 1, where $\otimes$ denotes the Kronecker product. To provide unbiased estimators of the respective target space, first order methods require that $E(Z|\beta^T Z)$ is linear in $\beta^T Z$, which is known as the linear conditional mean assumption. Second-order methods, in addition, require $\mathrm{Var}(Z|\beta^T Z)$ to be nonrandom, which is known as the constant conditional variance assumption.

Imposing such assumptions about the predictor is often considered as the necessary tradeoff for overcoming the curse of dimensionality until recent developments of central solution space [14,9], which provide sufficient dimension reduction estimators based on estimating equations. Li and Dong [14] defined the SIR-based central solution space as the intersection of $\mathrm{Span}(\beta)$ over all β satisfying

$$E(Z|Y) = E[E(Z|\beta^T Z)|Y] \quad \text{a.s.} \tag{1}$$

Denote the basis of this central solution space as $\beta_{\mathrm{CSS-SIR}}$. It was shown in [14] that $\mathrm{Span}(\beta_{\mathrm{CSS-SIR}}) \subseteq \mathcal{S}_{Y|Z}$. Under the linear conditional mean assumption, $\mathrm{Span}(\beta_{\mathrm{CSS-SIR}}) = \mathrm{Span}(\beta_{\mathrm{SIR}})$ and CSS-SIR reduces to the classical SIR estimator. Central solution space estimators are thus generalization of traditional moment-based estimators. Dong and Li [9] defined the SAVE-based central solution space as the intersection of $\mathrm{Span}(\beta)$ over all β such that

$$\mathrm{Var}(Z) - \mathrm{Var}(Z|Y) = \mathrm{Var}[E(Z|\beta^T Z)] - \mathrm{Var}[E(Z|\beta^T Z)|Y] \quad \text{a.s.} \tag{2}$$

However, estimators discussed in [14,9] are all targeting at the central space $\mathcal{S}_{Y|X}$. They did not address the cases when the target is the central mean space or the central kth moment space, which might be a proper subspace of the central space. Moreover, central solution space estimators based on (1) and (2) require slicing and need to specify the number of slices as a tuning parameter. While [25] proved that SIR is consistent with each slice containing a fixed number of data, SAVE is inconsistent under the same setting [17]. As a result, the SAVE estimator is sensitive to the choice of slice numbers. It has not been studied before whether the choice of slice numbers will affect the performances of slicing-based central solution space estimators.

This paper is organized as follows. In Section 2, we revisit existing estimators in the central kth moment space. In Section 3, two sets of estimating equations are introduced that hold without the linear conditional mean assumption. These equations are then leading to our new estimators. They also give rise to the notion of the central kth moment solution space which is also introduced and discussed in Section 3. One of our proposed estimators can be seen as a generalization of the covariance estimator introduced by Yin and Cook [23]. Section 4 discusses algorithms for computing our proposed estimators. Section 5 discusses how to determine the structural dimension d of the central kth moment space. To compare our proposed estimators with existing estimators, numerical studies are performed in Section 6 and a real data analysis is implemented in Section 7. Conclusions are stated in Section 8 followed by the Appendix with proofs.

2. Existing estimators of the CKMS

Let β be a basis of $\mathcal{S}_{Y|Z}^{(k)}$. Denote $P_\beta = \beta(\beta^T\beta)^{-1}\beta^T$ as the projection with respect to the inner product $\langle a, b\rangle = a^T b$ on to Span(β). The linear conditional mean assumption requires $E(Z|\beta^T Z)$ to be linear in $\beta^T Z$, which implies $E(Z|\beta^T Z) = P_\beta Z$. There are two families of CKMS estimators in the literature: the OLS-based and the PHD-based methods.

Denote $f(Y)$ as any polynomial of Y with degree no more than k, then $E[f(Y)|\beta^T Z] = E[f(Y)|Z]$ from the definition of $\mathcal{S}_{Y|Z}^{(k)}$. It follows that

$$E[Zf(Y)] = E\{ZE[f(Y)|Z]\} = E\{ZE[f(Y)|\beta^T Z]\} = E[E(Z|\beta^T Z)f(Y)].$$

Under the linear conditional mean assumption, we then have

$$E[Zf(Y)] = E[E(Z|\beta^T Z)f(Y)] = P_\beta E[Zf(Y)] \in \mathcal{S}_{Y|Z}^{(k)}. \tag{3}$$

Yin and Cook [23] chose $f(Y)$ as powers of Y up to order k, and proposed to estimate the CKMS by Span$\{E(YZ), \ldots, E(Y^k Z)\}$, which was denoted as the covariance subspace $\mathcal{S}_{\mathrm{COV}}^{(k)}$. The OLS estimator is a special case of COV with $k = 1$.

Some limitations of COV estimation are inherited from OLS estimation. In particular, a COV estimator can estimate at most k directions in $\mathcal{S}_{Y|Z}^{(k)}$. Also, it is not very effective in recovering $\mathcal{S}_{Y|Z}^{(k)}$ when the link function between the response and the predictors is U-shaped. The latter is a limitation shared by all the first-order methods in Table 1, and can be mitigated by the second-order estimators. Next we assume that $\mathrm{Var}(Z|\beta^T Z)$ is nonrandom, which is a common assumption among second-order methods. Denote $\tilde{f}(Y) = f(Y) - E[f(Y)]$ and we have

$$E[\tilde{f}(Y)ZZ^T] = E\{E[\tilde{f}(Y)|Z]ZZ^T\} = E[\tilde{f}(Y)E(ZZ^T|\beta^T Z)].$$

Plug in $E(ZZ^T|\beta^T Z) = \mathrm{Var}(Z|\beta^T Z) + E(Z|\beta^T Z)E^T(Z|\beta^T Z)$ and use the fact that $E[\tilde{f}(Y)\mathrm{Var}(Z|\beta^T Z)] = E[\tilde{f}(Y)]\mathrm{Var}(Z|\beta^T Z) = 0$, we get

$$E[\tilde{f}(Y)ZZ^T] = E[\tilde{f}(Y)E(Z|\beta^T Z)E^T(Z|\beta^T Z)] = P_\beta E[\tilde{f}(Y)ZZ^T]P_\beta \in \mathcal{S}_{Y|Z}^{(k)}.$$

This is essentially the estimator proposed by Yin and Bura [22]. They also relaxed the constant conditional variance assumption with a less restrictive one that requires $\mathrm{Var}(Z|\beta^T Z)$ to be uncorrelated with $\tilde{f}(Y)$. With $f(Y) = Y^k$ and $k = 1$, this reduces to the classical y-based PHD estimator [13]. It is worth mentioning that the estimator based on PHD can not recover $\mathcal{S}_{Y|Z}^{(k)}$ when the link function between the response and the predictors is linear.

3. Central *k*th moment solution space

In this section, we first examine the limitations of existing CKMS estimators, and then introduce new CKMS estimators without requiring the linearity assumption. The connections between our proposed estimators and existing estimators of the CKMS are also revealed.

3.1. Limitations of existing CKMS estimators

Both the OLS-based and PHD-based estimators in the CKMS require the linear conditional mean assumption: $E(Z|\beta^T Z)$ is linear in $\beta^T Z$, where β is the basis of $\mathcal{S}_{Y|Z}^{(k)}$. Since β is unknown in practice, the linearity assumption is made for all possible β, which implies the distribution of Z has to be elliptically-contoured [11]. When the linearity assumption is not met, existing estimators of CKMS will fail. We provide the following examples to fix the idea. X scale predictors are used for the ease of illustration.

Example 1. $Y = X_1 + X_1X_2 + \epsilon, X = (X_1, X_2, X_3)^T \sim N(0, I_3), \epsilon \sim N(0, 1)$ and $\epsilon \perp\!\!\!\perp X$. This example was examined by Yin and Cook [23]. OLS alone is not able to fully recover $\mathcal{S}_{Y|X}$ since $\beta_{\mathrm{OLS}} = \Sigma^{-1}E(XY) = (1, 0, 0)^T$. Yin and Cook [23] suggested including $\Sigma^{-1}E(XY^2) = (0, 2, 0)^T$ to exhaustively estimate $\mathcal{S}_{Y|X} = \mathcal{S}_{Y|X}^{(1)}$. Next, we introduce non-ellipticity in X and let $(X_1, X_2)^T \sim N(0, I_2), X_3 = X_1^2 + X_1 - 1 + \delta, \delta \sim N(0, 1)$ and $\delta \perp\!\!\!\perp (X_1, X_2, \epsilon)$. Then $\beta_{\mathrm{OLS}} = (1, 0, 0)^T$ is still unbiased, but $\Sigma^{-1}E(XY^2) = (-4/3, 2, 4/3)^T$ is biased.

Example 1 (*Continued*)**.** Keep the setup as before but assume that $Y = X_1 + 1 + (X_2 + 1)\epsilon$. With normally distributed predictors, simple calculations show $\beta_{\mathrm{OLS}} = (1, 0, 0)^T$ and $\Sigma^{-1}E(XY^2) = (2, 2, 0)^T$. Together they fully recover $\mathcal{S}_{Y|X} = \mathcal{S}_{Y|X}^{(2)}$. In the non-elliptical case, we still have $\beta_{\mathrm{OLS}} = (1, 0, 0)^T$, but $\Sigma^{-1}E(XY^2) = (4/3, 2, 2/3)^T$ is no longer in $\mathcal{S}_{Y|X}^{(2)}$.

Example 2. The setup is the same as in Example 1, but we assume the quadratic relation $Y = X_1^2 + X_2^2\epsilon$. With normally distributed predictors, $E(XY^k) = 0$ and OLS-type estimators fail for any integer k. On the other hand, $\Sigma^{-1}\text{Span}(E\{[Y - E(Y)]XX^T\}) = (1, 0, 0)^T$ and PHD only detects the direction in the central mean space. Yin and Bura [22] suggested using $\Sigma^{-1}\text{Span}(E\{[Y^2 - E(Y^2)]XX^T\}) = (1, 0, 0)^T \cup (0, 1, 0)^T$ and this yields an additional direction in the variance component. In the non-elliptical case, the estimators based on [22] no longer work as they become

$$\Sigma^{-1}E\{[Y - E(Y)]XX^T\} = \begin{pmatrix} 2 & 0 & -2/3 \\ 0 & 0 & 0 \\ 0 & 0 & 8/3 \end{pmatrix}$$

and

$$\Sigma^{-1}E\{[Y^2 - E(Y^2)]XX^T\} = \begin{pmatrix} 12 & 0 & -12 \\ 0 & 12 & 0 \\ 0 & 0 & 24 \end{pmatrix}.$$

3.2. New estimator based on OLS

Let $T_k = (Y, Y^2, \ldots, Y^k)$ as in Table 1. Eq. (3) implies

$$E(T_k \otimes Z) = E[T_k \otimes E(Z|\beta^T Z)] = E[T_k \otimes (P_\beta Z)] \in \mathcal{S}^k(Y|Z).$$

The first equality above is guaranteed by the conditional independency in Definition 1 of CKMS, and only the second equality above requires the linearity assumption. This motivates the following definition.

Definition 2. The OLS-based central kth moment solution space is defined to be $\mathcal{S}_{\text{CKMSS-OLS}} = \cap\text{Span}(\beta)$, where the intersection is over all β that satisfies equation

$$E(T_k \otimes Z) = E[T_k \otimes E(Z|\beta^T Z)] \quad \text{a.s.} \tag{4}$$

Estimators of the CKMS based on (4) do not rely on the restrictive linearity assumption, as we will see next.

Proposition 1. *Suppose all the moments involved are finite. Then the following relations hold.*

1. $\mathcal{S}_{\text{CKMSS-OLS}} \subseteq \mathcal{S}_{Y|Z}^{(k)}$.
2. *Let β be a basis of $\mathcal{S}_{Y|Z}^{(k)}$. If $E(Z|\beta^T Z)$ is a linear function of $\beta^T Z$, then* $\mathcal{S}_{\text{CKMSS-OLS}} = \mathcal{S}_{\text{COV}}^{(k)} \subseteq \mathcal{S}_{Y|Z}^{(k)}$.

Proposition 1 is parallel to Theorem 2.1 of Li and Dong [14]. Part 1 says that $\mathcal{S}_{\text{CKMSS-OLS}}$ belongs to the CKMS with essentially no assumptions other than the finiteness of moments. When the linear conditional mean assumption holds, $\mathcal{S}_{\text{CKMSS-OLS}}$ reduces to the covariance subspace $\mathcal{S}_{\text{COV}}^{(k)}$.

As a special case with $k = 1$, Eq. (4) becomes

$$E(ZY) = E[E(Z|\beta^T Z)Y] \quad \text{a.s.} \tag{5}$$

The intersection of $\text{Span}(\beta)$ over all β satisfying Eq. (5) is defined to be the OLS-based central mean solution space, and denoted by $\mathcal{S}_{\text{CMSS-OLS}}$. Li and Dong [14] suggested that this quantity belongs to the $\mathcal{S}_{Y|Z}$. We know more precisely that $\mathcal{S}_{\text{CMSS-OLS}} \subseteq \mathcal{S}_{E(Y|Z)} \subseteq \mathcal{S}_{Y|Z}$.

3.3. New estimator based on PHD

Denote $T_k = (Y, Y^2, \ldots, Y^k)$ as before. We have

Definition 3. The PHD-based central kth moment solution space is defined to be $\mathcal{S}_{\text{CKMSS-PHD}} = \cap\text{Span}(\beta)$, where the intersection is over all β that satisfies equation

$$E\{[T_k - E(T_k)] \otimes (ZZ^T)\} = E\{[T_k - E(T_k)] \otimes [E(Z|\beta^T Z)E^T(Z|\beta^T Z)]\} \quad \text{a.s.} \tag{6}$$

The next proposition states the relationship between $\mathcal{S}_{\text{CKMSS-PHD}}$ and $\mathcal{S}_{Y|Z}^{(k)}$.

Proposition 2. *Suppose $Y^j - E(Y^j)$ is uncorrelated with* $\text{Var}(Z|\beta^T Z)$ *for $j = 1, \ldots, k$, then*

1. $\mathcal{S}_{\text{CKMSS-PHD}} \subseteq \mathcal{S}_{Y|Z}^{(k)}$.
2. *Let β be a basis of $\mathcal{S}_{Y|Z}^{(k)}$. If $E(Z|\beta^T Z)$ is a linear function of $\beta^T Z$, then* $\mathcal{S}_{\text{CKMSS-PHD}} = \text{Span}(E\{[T_k - E(T_k)] \otimes (ZZ^T)\}) \subseteq \mathcal{S}_{Y|Z}^{(k)}$.

As a special case with $k = 1$, Eq. (6) becomes

$$E\{[Y - E(Y)]ZZ^T\} = E\{[Y - E(Y)]E(Z|\beta^T Z)E^T(Z|\beta^T Z)\}. \tag{7}$$

The intersection of Span(β) over all β satisfying Eq. (7) is defined as the PHD-based central mean solution space, and denoted by $\mathcal{S}_{\mathrm{CMSS-PHD}}$.

Among estimators of the central mean space, we have seen in examples from Section 2 that OLS cannot recover patterns which are symmetric about the origin, while PHD will fail when the link function is linear. The extensions we make in Section 3 aim at providing estimators in $\mathcal{S}_{E(Y|Z)}$ and $\mathcal{S}^{(k)}_{Y|Z}$ that do not require the linear conditional mean assumption. Our proposed estimators will work for non-elliptically distributed predictors, but they inherit their respective limitations from OLS and PHD. More specifically, when a linear trend is revealed between the response and the predictor, then OLS-based methods should be used to estimate $\mathcal{S}_{E(Y|Z)}$ or $\mathcal{S}^{(k)}_{Y|Z}$; when the preliminary analysis or visualization suggests a symmetric link function, then PHD-based methods may be preferred.

4. Estimating the central *k*th moment solution space

To find solutions to estimating equations at the sample level, we suggest two different approaches. The first approach is along the lines of Li and Dong [14], and we replace solving estimation equations with an equivalent optimization problem. Reliable numerical minimization algorithms play an important role in finding accurate estimators for this approach. The second approach utilizes the connection between equation-based estimators and classical estimators revealed in Propositions 1 and 2, and suggests an iterative algorithm that alternates between OLS regression and eigenvalue decomposition. The second approach does not rely on numerical optimization and is suitable for applications with large sample size and high dimension predictor.

Our discussion in this section will be based on Eq. (7). Estimating the solution spaces for Eqs. (4)–(6) in Section 3 can be implemented in a parallel fashion. Without loss of generality, assume $E(Y) = 0$ in (7) throughout this section.

4.1. Minimizing objective function approach

Assume the structural dimension d is known. Borrowing the idea of [14], solving Eq. (7) can be transformed to minimizing over $\beta \in \mathbb{R}^{p\times d}$ the following objective function

$$L(\beta) = \|E\{Y[ZZ^T - E(Z|\beta^T Z)E^T(Z|\beta^T Z)]\}\|^2,$$

where $\|\cdot\|$ denotes the Frobenious matrix norm.

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be independent and identically distributed copies of (X, Y). For a function $r(X, Y)$, denote $E_n[r(X, Y)]$ as the sample average $n^{-1}\sum_{i=1}^n r(X_i, Y_i)$. The sample level objective function becomes

$$L_n(\beta) = \|E_n\{Y[ZZ^T - g(\beta^T Z)g^T(\beta^T Z)]\}\|^2, \tag{8}$$

where $g(\beta^T Z)$ is the sample estimate of $E(Z|\beta^T Z)$. Choose a set of basis functions that are sufficiently flexible to describe $E(Z|\beta^T Z)$, and denote it as $G(\beta^T Z) = \left\{f_1(\beta^T Z), \ldots, f_s(\beta^T Z)\right\}^T$. Then we have

$$g(\beta^T Z) = E_n(Z|\beta^T Z) = E_n[ZG^T(\beta^T Z)]\{E_n[G(\beta^T Z)G^T(\beta^T Z)]\}^{-1}G(\beta^T Z).$$

If we know the true underlying function form of $E(Z|\beta^T Z)$ and use the exact basis for $G(\beta^T Z)$, we will have unbiased and consistent estimator of the CKMS. In the presence of apparent nonlinear relationship among the predictors, we can still get better estimation compared with classical estimators, as we model such nonlinearity rather than assuming the relation to be linear. In our simulation, we use cubic polynomial functions of $\beta_i^T Z$ as the basis, where β_i's are columns of β for $i = 1, \ldots, d$. Our experiences indicate that many sensible choices of $G(\beta^T Z)$ can improve upon classical CKMS estimators with non-elliptically distributed predictors. Cross-validation can be used to evaluate different choices of basis functions in practice, as we will see in the real data application in Section 7.

Note that Span(β), instead of β itself, is the true parameter of interest in (8). Li and Dong [14] suggest to use the R-function *OPTIM* that utilizes the Nelder–Mead simplex method [19] for the optimization. Another possibility is to implement specialized gradient descent algorithm that applies to optimization over the Grassmann manifold $\mathcal{G}_{d,p}$ [18,6]. Details about implementation of this algorithm can be found in [10]. As suggested in [14], we use the outer product gradient estimators (OPG) [20] as the initial value, which does not rely on the linearity assumption but involves high-dimensional smoothing. Our experience indicates OPG works well as the initial value for small to moderate predictor dimension p.

4.2. Iterative alternating algorithm approach

Recently, Li and Dong [15] developed an iterative algorithm to deal with non-elliptical predictors that target at the central space. We now adapt their procedure for our CKMSS estimators.

Assume that the basis functions in $G(\beta^T Z)$ include, but are not limited to the linear functions $\beta_1^T Z, \ldots, \beta_d^T Z$. Let $\delta_1(\beta^T Z), \ldots, \delta_{s-d}(\beta^T Z)$ denote functions in $G(\beta^T Z)$ other than the linear ones, and let $\delta(\beta^T Z)$ be the $s-d$ dimensional vector consisting of these functions. Then, there are non-random matrices $A_1 \in \mathbb{R}^{p\times d}$ and $A_2 \in \mathbb{R}^{p\times(s-d)}$ such that

$$E(Z|\beta^T Z) = A_1\beta^T Z + A_2\delta(\beta^T Z). \tag{9}$$

With $\delta(\beta^T Z) = 0$, model (9) reduces to the traditional linear conditional mean assumption. The next proposition is the basis for our new algorithm.

Proposition 3. *Under model (9), Eq. (7) is equivalent to*

$$E(YZ^*Z^{*T}) = E[YE(Z^*|\beta^T Z^*)E^T(Z^*|\beta^T Z^*)] \quad a.s. \tag{10}$$

where $Z^* = Z - A_2\delta(\beta^T Z)$. Furthermore, $E(Z^*|\beta^T Z^*)$ is linear in $\beta^T Z^*$.

Because Z^* satisfies the linearity assumption, by Proposition 2, $\beta \in \mathbb{R}^{p\times d}$ that solves Eq. (10) will guarantee $\mathrm{Span}(\beta)$ coincides with $\mathrm{Span}\{E(YZ^*Z^{*T})\}$, the column space of the classical y-based PHD estimator. This implies that given β, we can update Z^* based on model (9); given Z^*, we can perform y-based PHD algorithm [13] to update β. We now propose an alternating algorithm.

1. Center Y_i as $\tilde{Y}_i = Y_i - E_n(Y)$ and standardize X_i as $\tilde{Z}_i = \hat{\Sigma}^{-1}[X_i - E_n(X)]$, where $\hat{\Sigma} = E_n\{[X - E_n(X)][X - E_n(X)]^T\}$. Choose basis functions $f_1(\beta^T Z), \ldots, f_s(\beta^T Z)$, which contain $\beta_1^T Z, \ldots, \beta_d^T Z$. β_i denotes the ith column of β.
2. Set initial value of β to be $\hat{\beta}_{(0)}$.
3. At the lth iteration, let $\hat{\beta}_{(l)}$ be the estimate of β. Perform linear regression of $\tilde{Z}_i$ on $\{f_1(\hat{\beta}_{(l)}^T\tilde{Z}_i), \ldots, f_s(\hat{\beta}_{(l)}^T\tilde{Z}_i)\}$. Denote

   $$A_1\hat{\beta}_{(l)}^T\tilde{Z}_i + A_2\delta(\hat{\beta}_{(l)}^T\tilde{Z}_i) = AV_i,$$

   where $A = (A_1, A_2)$ and

   $$V_i = \begin{pmatrix} \hat{\beta}_{(l)}^T\tilde{Z}_i \\ \delta(\hat{\beta}_{(l)}^T\tilde{Z}_i) \end{pmatrix}.$$

   Then the least squares estimate of A is

   $$\hat{A} = E_n(\tilde{Z}V^T)[E_n(VV^T)]^{-1}.$$

   Let $\hat{A}_1$ be the first d columns of $\hat{A}$, and $\hat{A}_2$ be the remaining columns.
4. Let $\tilde{Z}_{i(l)}^* = \tilde{Z}_i - \hat{A}_2\delta(\hat{\beta}_{(l)}^T\tilde{Z}_i)$. Then perform the y-based PHD to obtain $\hat{\beta}_{(l+1)}$ from the adjusted sample $(\tilde{Z}_{i(l)}^*, Y_i)$.
5. Stop the iteration after either (a) the lth iteration if a convergence rule is met, or (b) a fixed maximum number. Otherwise go back to step 3.

We use OPG as the initial value unless specified otherwise. To measure the closeness between consecutive estimators, we use $\hat{r}^2(\hat{\beta}_{(l)}^T\tilde{Z}, \hat{\beta}_{(l-1)}^T\tilde{Z})$, the sample squared trace correlation between $\hat{\beta}_{(l)}^T\tilde{Z}$ and $\hat{\beta}_{(l-1)}^T\tilde{Z}$. We stop the iteration if this correlation becomes greater than a certain threshold, say 0.99. For random vectors $U, V \in \mathbb{R}^d$, denote $t_1, \ldots, t_v$ as the nonzero eigenvalues of $\{\mathrm{Var}(U)\}^{-1/2}\mathrm{Cov}(U, V)\{\mathrm{Var}(V)\}^{-1}\mathrm{Cov}(V, U)\{\mathrm{Var}(U)\}^{-1/2}$. The squared trace correlation between U and V is

$$r^2(U, V) = d^{-1}\sum_{i=1}^{v} t_i. \tag{11}$$

We set maximum number of iterations to be 20 in our simulation.

5. Order determination

We have so far assumed that the true structural dimension d of the CKMS is known. A sequential testing approach is commonly used to determine unknown structural dimension. Denote ℓ as the working structural dimension. Then test $H_0 : d = \ell$ versus $H_a : d > \ell$ for $\ell = 0, 1, \ldots, p$. The estimate $\hat{d}$ of the structural dimension is such that H_0 is rejected for $\ell = 0, \ldots, \hat{d} - 1$, and H_0 is not rejected for $\ell = \hat{d}$.

Chi-squared tests, weighted Chi-squared tests, and permutation tests are commonly used for testing the rank of a given dimension reduction kernel matrix. For example, the y-based PHD uses the kernel matrix $\Sigma_{yzz} = E\{[Y - E(Y)]ZZ^T\}$. Let $|\hat{\lambda}_1| \geq |\hat{\lambda}_2| \geq \cdots \geq |\hat{\lambda}_p|$ be the ordered absolute eigenvalues of $\hat{\Sigma}_{yzz}$, the sample estimate of Σ_{yzz}. The corresponding test statistic is

$$\Lambda_\ell = \frac{n}{2\widehat{\mathrm{Var}}(Y)}\sum_{i=\ell+1}^{p}\hat{\lambda}_i^2,$$

where $\widehat{\mathrm{Var}}(Y)$ is the sample estimate of $\mathrm{Var}(Y)$. For normally distributed predictors, Li [13] showed that Λ_ℓ follows an asymptotic Chi-squared distribution with $(p-\ell)(p-\ell+1)/2$ degrees of freedom under $H_0 : d = \ell$. Cook [3] revisited this and proved in the more general case with elliptically distributed predictors, that Λ_ℓ has a weighted Chi-squared asymptotic distribution under the null. As an alternative, Cook and Yin [8] suggested the permutation test for order determination, where the observed test statistic Λ_ℓ is compared with its permutation distribution under the null hypothesis. As we will see in Section 6, all three tests will fail to consistently estimate the true structural dimension with non-elliptically distributed predictors.

Asymptotic tests for sufficient dimension reduction estimators based on estimating equations are not available in the literature. As an alternative, we use bootstrap for order determination. This idea was first introduced into the dimension reduction literature by Ye and Weiss [21], and has been implemented by Dong and Li [9] for determining the structural dimension of the central solution space. For a working structural dimension ℓ, let $\hat{\beta}_{\ell,0}$ be the estimate from the original sample. Let $\hat{\beta}_{\ell,b}$, $b = 1, \ldots, B$ be the estimates based on the bth bootstrap sample. We then estimate d by maximizing, over $\ell = 1, \ldots, p-1$, the following quantity,

$$\bar{r}_\ell^2 = \frac{1}{B}\sum_{b=1}^{B} \hat{r}^2(\hat{\beta}_{\ell,b}^T X, \hat{\beta}_{\ell,0}^T X), \tag{12}$$

where $\hat{r}^2(\hat{\beta}_{\ell,b}^T X, \hat{\beta}_{\ell,0}^T X)$ is the sample estimate of the square trace correlation defined by (11). This criteria is similar in spirit to those used in [21], but we take into account the variation of X.

6. Numerical studies

First we study the following four models: (I) $Y = \exp(X_1) + 0.2\epsilon$, (II) $Y = X_1^2 + 0.2\epsilon$, (III) $Y = X_1 + X_1X_2 + 0.2\epsilon$, and (IV) $Y = X_1 + 1 + (X_2 + 1)\epsilon$. Here ϵ is standard normal error independent of X. We consider two settings for X: (i) $X = (X_1, X_2, \ldots, X_p)^T$ has standard multivariate normal distribution; (ii) $(X_1, X_2)^T \sim N(0, I_2)$, $X_3 = X_1^2 + X_1 - 1 + \delta$ with $\delta \sim N(0, 1)$ and $\delta \perp\!\!\!\perp (X_1, X_2, \epsilon)$. For $j > 3$, $X_j \perp\!\!\!\perp (X_1, X_2, X_3, \epsilon)$ are taken to be independent $N(0, 1)$. X is normal in case (i) and non-elliptical in case (ii).

To compare the performance of different estimators, we use the sample squared trace correlation $\hat{r}^2 = \hat{r}^2(\beta^T X, \hat{\beta}^T X)$, whose population version is defined by (11). The closer the correlation $\hat{r}^2$ is to 1, the better the estimator is. Each entry of Tables 2–5 is formatted as $a(b)$, where a is the average of the correlation across the 100 simulated samples, and b is the standard error of this average.

We summarize the results for Model I in Table 2. Under both the elliptical and the non-elliptical distributions of X, we compare OPG and classical OLS estimators with our proposed estimators from solving Eq. (5). Denote the optimization-based estimator from Section 4.1 as CMSS-OLS and the estimator from Section 4.2 as ITE-OLS. CSS-SIR estimators are also included for a complete comparison. Please refer to [14] for the algorithm of CSS-SIR. CSS-SIR requires slicing the range of Y and we denote the slice number by h. Because of the monotone trend in the link function, the second-order estimators are not as effective in this case and thus not reported.

From Table 2, we see that when X is elliptical in case (i), all methods work very well. OPG is worse than the other methods because it involves high-dimensional smoothing. When X is non-elliptical in case (ii), the linear conditional mean assumption is no longer satisfied, and classical OLS becomes the worst. ITE-OLS is only slightly worse than CMSS-OLS, which performs the best among all estimators. Although its performance deteriorates when X is non-elliptical with slice number $h = 2$, CSS-SIR appears to be not sensitive to h in general. This result resonates with the finding in [25], where classical SIR is shown to be insensitive to the choice of slice numbers.

The results for Model II are reported in Table 3. OLS-type methods do not work in this case as the response is symmetric about X. We compare OPG and classical y-based PHD with our proposed estimators from solving Eq. (7). Denote the estimator from Section 4.1 as CMSS-PHD and the estimator from Section 4.2 as ITE-PHD. We also include CSS-SAVE estimators based on solving Eq. (2) across different slice numbers $h = 2, 10, 20$. Details about estimating CSS-SAVE are provided in [9]. The trend in Table 3 is similar to what we have observed in Table 2. All three PHD-based methods work very well when X is normal. In case (ii) when X is non-elliptical, classical PHD will fail. CMSS-PHD has the best performance at the expense of heavier computation. ITE-PHD is much faster than CKSS-PHD but less accurate. The accuracy of CSS-SAVE estimators may vary a lot for different choices of h. Li and Zhu [17] demonstrated that the SAVE estimator is sensitive to the choice of slice numbers, and our results show that CSS-SAVE inherits this disadvantage.

In Table 4 we consider Model III and Model IV with fixed sample size $n = 200$. The classical COV is compared with the iterative COV (ITE-COV), which is based on the iterative alternating algorithm introduced in Section 4.2. Model IV is heteroscedastic, and its direction in the variance component can not be recovered by OPG, which only targets the central mean space. Hence the classical COV estimator is used as the initial value of ITE-COV instead. These two models are modified from Example 1, where we have seen that the classical COV estimator fully recovers the central space when X is normal, and becomes biased with non-elliptical X. We observe here that the ITE-COV estimator consistently improves the classical COV for non-elliptical X. Not surprisingly, all estimators become worse as the predictor dimension p increases.

Table 2
Model I with $n = 200$ and $p = 8$. Based on 100 repetitions, average of the sample squared trace correlation $\hat{r}^2(\beta^T X, \hat{\beta}^T X)$ and its standard error are reported.

X	OLS	OPG	CMSS-OLS	ITE-OLS	CSS-SIR		
					$h = 2$	$h = 10$	$h = 20$
(i)	.980 (.001)	.958 (.003)	.999 (.0001)	.999 (.0001)	.980 (.001)	.993 (.0003)	.992 (.0003)
(ii)	.767 (.004)	.947 (.003)	.998 (.0002)	.985 (.002)	.951 (.002)	.991 (.0005)	.989 (.0005)

Table 3
Model II with $n = 200$ and $p = 8$. Based on 100 repetitions, average of the sample squared trace correlation $\hat{r}^2(\beta^T X, \hat{\beta}^T X)$ and its standard error are reported.

X	PHD	OPG	CMSS-PHD	ITE-PHD	CSS-SAVE		
					$h = 2$	$h = 10$	$h = 20$
(i)	.949 (.002)	.956 (.002)	.997 (.0002)	.994 (.001)	.942 (.002)	.824 (.011)	.805 (.011)
(ii)	.353 (.017)	.940 (.004)	.960 (.006)	.913 (.020)	.950 (.005)	.935 (.007)	.909 (.010)

Table 4
Non-elliptical X in case (ii) with $n = 200$. Based on 100 repetitions, average of the sample squared trace correlation $\hat{r}^2(\beta^T X, \hat{\beta}^T X)$ and its standard error are reported.

Model	III		IV	
p	COV	ITE-COV	COV	ITE-COV
4	.728 (.005)	.916 (.007)	.846 (.006)	.930 (.004)
6	.714 (.005)	.896 (.007)	.806 (.006)	.892 (.005)
8	.699 (.004)	.901 (.007)	.799 (.005)	.868 (.004)
10	.698 (.004)	.885 (.006)	.781 (.005)	.835 (.004)

Table 5
Non-elliptical X in case (ii) with $p = 20$. Based on 100 repetitions, average of the sample squared trace correlation $\hat{r}^2(\beta^T X, \hat{\beta}^T X)$ and its standard error are reported.

Model	III		IV	
n	COV	ITE-COV	COV	ITE-COV
400	.682 (.002)	.886 (.004)	.762 (.003)	.827 (.002)
600	.701 (.002)	.904 (.003)	.796 (.002)	.879 (.001)
800	.697 (.001)	.916 (.002)	.813 (.002)	.906 (.001)
1000	.702 (.001)	.920 (.002)	.831 (.001)	.924 (.001)

Next we consider $p = 20$ with $n = 400, 600, 800, 1000$ and summarize the results in Table 5. With large sample size and high predictor dimension in this setting, the optimization-based estimator from Section 4.1 is not desirable as it is very intensive computationally. ITE-COV with classical COV as its initial value is appealing for the ease of computation. Even with non-elliptical X, large predictor dimension, and an initial value that is less than ideal, our proposed ITE-COV estimator still has good overall performance.

Now we turn to order determination. OLS can estimate at most one direction in the central mean space. Similarly, COV can estimate at most k directions in the CKMS. In practice, we are mostly concerned with $E(Y|X)$ and $\mathrm{Var}(Y|X)$ and k is often set to be 2 for COV estimators. Thus our discussion of order determination will focus on PHD-based tests. The models we consider are one-dimensional Model (I) $Y = \exp(X_1) + 0.2\epsilon$ and two-dimensional Model (III) $Y = X_1 + X_1X_2 + 0.2\epsilon$. We compare the Chi-squared test, weighted Chi-squared test, permutation test, and bootstrap. Set $p = 8$, $n = 400$, and the proportions of the estimated structural dimension $\hat{d}$ is reported over 100 repetitions.

From Table 6, we see that when X is normal in case (i), all four tests can correctly specify the underlying structural dimension d most of the times for both Model I and Model III. When X is non-elliptical in case (ii), all four tests still seem to work well for Model III and correctly specify $\hat{d} = 2$ with a decent proportion. However, for Model I which is one-dimensional, the Chi-squared test, weighted Chi-squared test and permutation test will overestimate the structural dimension to be $\hat{d} = 2$ with a large proportion, and only the bootstrap test correctly specify the true structural dimension 91 times out of 100 repetitions. Thus we conclude bootstrap enjoys the best overall performance for order determination.

Table 6
Order determination with $n = 400$ and $p = 8$. Proportions based on 100 repetitions are reported.

X	Model	Test	$\hat{d} = 0$	$\hat{d} = 1$	$\hat{d} = 2$	$\hat{d} > 2$
(i)	I	χ^2	0	**.97**	.03	0
		Weighted χ^2	.38	**.61**	01	0
		Permutation	0	**.96**	.03	.01
		Bootstrap	NA	**.87**	0	.13
	III	χ^2	0	0	**.95**	.05
		Weighted χ^2	0	.01	**.99**	0
		Permutation	0	0	**.91**	.09
		Bootstrap	NA	.19	**.80**	.01
(ii)	I	χ^2	0	.26	**.65**	.09
		Weighted χ^2	.16	.17	**.67**	0
		Permutation	0	.25	**.65**	.01
		Bootstrap	NA	**.91**	0	09
	III	χ^2	0	0	**.89**	.11
		Weighted χ^2	.03	.04	**.93**	0
		Permutation	0	0	**.86**	.14
		Bootstrap	NA	.11	**.83**	.06

Table 7
Order determination for the Massachusetts college data. Based on 200 bootstrap samples, $\bar{r}_\ell^2$ in (12) is reported.

ℓ	1	2	3	4	5	6	7
$\bar{r}_\ell^2$	**.919**	.615	.638	.689	.721	.810	.899

Table 8
Comparison of different sufficient dimension reduction estimators of β for the Massachusetts college data. PRESS is reported based on leave-one-out cross validation.

CMSS-OLS			OPG	CSS-SIR			CSS-PIR
Linear	Quadratic	Cubic		$h = 2$	$h = 5$	$h = 10$	
4243	**3737**	4287	4251	6199	6965	6683	4345[a]

[a] Result from [14], where second-order polynomials were used for CSS-PIR.

7. Application to the Massachusetts college data

For an empirical study, we analyze the 1995 Massachusetts college data set, which is a built-in data set from MINITAB (release 15) and has been studied in Li and Dong [14]. We are interested in the percentage of freshmen that graduate at 46 different colleges. This percentage is our response Y with variable name "Grad". The predictor variables we use are "top25", "MSAT", "VSAT", "accept", "enroll", "tuition", "SFratio", "PubPriv", which are all variables that measure the quality of the incoming students or the features of the college. Please refer to MINITAB data description for the exact meaning of each variable. We restrict our attention to those variables for easy comparison with the result reported in [14]. The first 7 variables are continuous and the last one is binary. Suppose we are interested only in the mean level of the graduation percentage, and inference about $\mathcal{S}_{E(Y|X)}$ becomes our focus.

First we determine the structural dimension of $\mathcal{S}_{E(Y|X)}$. From the OLS sufficient plot in the left panel of Fig. 1, the link function for the regression mean seems to be linear, and PHD-based methods will not be suitable in this case. On the other hand, we observe a significant nonlinear relationship among some predictors in this data set. As we can see from the scatter plot in the right panel of Fig. 1, schools that charge very high or very low out-of-state tuition have a lower acceptance rate than schools with moderate tuition level. The Chi-squared test, weighted Chi-squared test, and permutation test all yield $\hat{d} = 2$. From the discussions in Section 6, this is likely to be an overestimate. Because OLS-based methods can estimate at most one direction in $\mathcal{S}_{E(Y|X)}$, we implement bootstrap with the OPG estimator. The results based on 200 bootstrap samples are summarized in Table 7. Because $\bar{r}_\ell^2$ is maximized at $\ell = 1$, we estimate $\hat{d} = 1$.

To compare performances of different dimension reduction estimators, we use leave-one-out cross validation. Let $\hat{\beta}_{-k}$ be the estimated β after deleting the kth observation (X_k, Y_k) from the sample. Then we fit a simple linear regression between Y_k and $\hat{\beta}_{-k}^T X_k$ and report the prediction error sum of squares (PRESS). We compare CMSS-OLS estimators from solving Eq. (5) and CSS-SIR estimators that solve Eq. (1). The last column in Table 8 is the original result from [14]. CSS-PIR is a variation of CSS-SIR that uses parametric inverse regression instead of slicing. For a fair comparison, all estimators are calculated by the minimizing objective function approach described in Section 4.1. To study the effect of using different

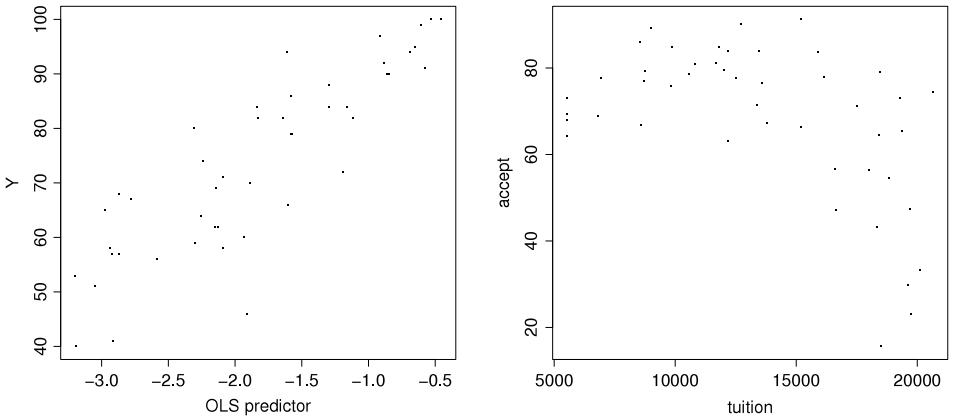

Fig. 1. OLS sufficient plot and scatter plot for the Massachusetts college data.

basis functions $G(\beta^T X)$, we consider linear, quadratic, and cubic polynomial functions of $\beta^T X$ for CMSS-OLS estimators. For CSS-SIR, we use quadratic polynomial functions as the basis, and consider slice numbers $h = 2, 5, 10$.

From Table 8, we see that CSS-SIR estimators have the worst performances. Due to the limited sample size $n = 46$, estimation based on even fewer intraslice observations does a poor job. By ignoring the variance component and focusing on the major trend in the mean component, CMSS-OLS and OPG estimators perform better than CSS-SIR and CSS-PIR estimators, which target the entire central space. CMSS-OLS with linear basis is equivalent to the classical OLS estimator, and it does a good job as we have seen in the left panel of Fig. 1. Because of the nonlinearity among the predictors, CMSS-OLS with quadratic basis improves over OLS and has the best performance. CMSS-OLS with cubic basis and OPG may have suffered from over-parametrization and high-dimensional smoothing respectively.

8. Conclusions

In this paper, we relax the linearity assumption required by existing estimators of the central mean space and the central kth moment space. For the sample level estimation, an iterative alternating algorithm is proposed for fast computation, which provides a desirable alternative to the optimization approach with large sample size and high predictor dimension. The linearity assumption is likely to fail when some or all of the predictors are discrete. Although the focus is on continuous predictors in this paper, our preliminary simulation study suggests that the central solution space framework has the potential to handle discrete predictors. A bootstrap methodology is recommended to choose the structural dimension, as the Chi-squared and weighted Chi-squared tests do not work as well for non-elliptical predictors. Order determination based on asymptotic tests for the proposed estimators requires further investigation.

Acknowledgments

This research was supported in part by the U.S. National Science Foundation Grant DMS-1106577 awarded to the first author. We sincerely thank an associate editor and three anonymous referees for giving useful comments that led to a much-improved presentation of the paper.

Appendix

Proof of Proposition 1. Part 1. Let $\text{Span}(\beta) = \mathcal{S}^{(k)}_{Y|Z}$. Then β satisfies

$$E(T_k \otimes Z) = E[E(T_k \otimes Z|Z)] = E[E(T_k|Z) \otimes Z] = E[E(T_k|\beta^T Z) \otimes Z].$$

The last equality above follows from the definition of $\text{Span}(\beta) = \mathcal{S}^{(k)}_{Y|Z}$. Then

$$E(T_k \otimes Z) = E[E(T_k|\beta^T Z) \otimes Z] = E[T_k \otimes E(Z|\beta^T Z)],$$

which means β satisfies Eq. (4). By the definition of $\mathcal{S}_{\text{CKMSS-OLS}}$, we have $\mathcal{S}_{\text{CKMSS-OLS}} \subseteq \text{Span}(\beta) = \mathcal{S}^{(k)}_{Y|Z}$. □

The proof of Part 2 needs the following Lemmas.

Lemma 1. *Let $f(Y)$ be any at most kth degree polynomial of Y and let $T_k = (Y, Y^2, \ldots, Y^k)$, then if β satisfies any one of the following three equations, it will also satisfy the other two:*

1. $E[Zf(Y)] = E[E(Z|\beta^T Z)f(Y)]$ *for any* $f(Y)$;
2. $E(ZY^i) = E[E(Z|\beta^T Z)Y^i]$ *for* $i = 1, \ldots, k$;
3. $E(T_k \otimes Z) = E[T_k \otimes E(Z|\beta^T Z)]$.

The proof of Lemma 1 is obvious and omitted.

Lemma 2. *Let* β *be a basis of* $\mathcal{S}_{\mathrm{COV}}^{(k)}$ *and assume* $E(Z|\beta^T Z)$ *is a linear function of* $\beta^T Z$. *Then* $\mathcal{S}_{\mathrm{COV}}^{(k)} = \mathrm{Span}(\gamma_1, \ldots, \gamma_k)$, *where* γ_i *solves equation* $E(ZY^i) = E[E(Z|\alpha^T Z)Y^i]$ *over* $\alpha \in \mathbb{R}^p$ *for* $i = 1, \ldots, k$.

Proof of Lemma 2. On one hand, we want to show for $i = 1, \ldots, k$, $\gamma_i \subseteq \mathcal{S}_{\mathrm{COV}}^{(k)} = \mathrm{Span}\{E(YZ), \ldots, E(Y^k Z)\}$. Since β is the basis for $\mathcal{S}_{\mathrm{COV}}^{(k)}$,

$$E(ZY^i) = P_\beta E(ZY^i) = E[(P_\beta Z)Y^i] = E[E(Z|\beta^T Z)Y^i].$$

The last equality above is guaranteed from the linearity assumption. Thus β solves $E(ZY^i) = E[E(Z|\alpha^T Z)Y^i]$. Because $\gamma_i \in \mathbb{R}^p$ also solves this equation with the smallest possible column space, we must have $\gamma_i \subseteq \mathrm{Span}(\beta) = \mathcal{S}_{\mathrm{COV}}^{(k)}$.

On the other hand, for $i = 1, \ldots, k$, γ_i satisfies

$$E(ZY^i) = E[E(Z|\gamma_i^T Z)Y^i] = P_{\gamma_i} E(ZY^i) = \gamma_i(\gamma_i^T \gamma_i)^{-1}\gamma_i^T E(ZY^i),$$

where the second equality follows from the linearity assumption. Thus $\mathcal{S}_{\mathrm{COV}}^{(k)} = \mathrm{Span}\{E(YZ), \ldots, E(Y^k Z)\} \subseteq \mathrm{Span}(\gamma_1, \ldots, \gamma_k)$. □

Proof of Proposition 1. Part 2. From Lemma 2, all we need to show now is $\mathcal{S}_{\mathrm{CKMSS\text{–}OLS}} = \mathrm{Span}(\gamma_1, \ldots, \gamma_k)$. Because $\gamma_i \in \mathbb{R}^p$ has the smallest column space among all possible β that satisfies $E(ZY^i) = E[E(Z|\beta^T Z)Y^i]$, $(\gamma_1, \ldots, \gamma_k)$ has the smallest column space among all possible β that simultaneously satisfies $E(ZY^i) = E[E(Z|\beta^T Z)Y^i]$ for $i = 1, \ldots, k$. From Lemma 1, this implies $(\gamma_1, \ldots, \gamma_k)$ has the smallest column space among all possible β that satisfies $E(T_k \otimes Z) = E[T_k \otimes E(Z|\beta^T Z)]$. By Definition 2, $\mathcal{S}_{\mathrm{CKMSS\text{–}OLS}}$ is unique and thus we have $\mathcal{S}_{\mathrm{CKMSS\text{–}OLS}} = \mathrm{Span}(\gamma_1, \ldots, \gamma_k)$. □

Proof of Proposition 2. Similar to proof of Proposition 1 and omitted. □

Proof of Proposition 3. We start from

$$Z = Z^* + A_2\delta(\beta^T Z). \tag{A.1}$$

Denote $U_\beta = E(Z|\beta^T Z)$ and $U_\beta^* = E(Z^*|\beta^T Z^*)$. Due to the fact that $E(Z^*|\beta^T Z) = E(Z^*|\beta^T Z^*)$ under model (9), take conditional expectation $E(\cdot|\beta^T Z)$ on both sides of (A.1) and we have

$$U_\beta = E(Z^*|\beta^T Z) + A_2\delta(\beta^T Z) = U_\beta^* + A_2\delta(\beta^T Z).$$

It follows that

$$\begin{aligned} E(YU_\beta U_\beta^T) &= E\{Y[U_\beta^* + A_2\delta(\beta^T Z)][U_\beta^* + A_2\delta(\beta^T Z)]^T\} \\ &= E(YU_\beta^* U_\beta^{*T}) + E[YU_\beta^*\delta^T(\beta^T Z)A_2^T] + E[YA_2\delta(\beta^T Z)U_\beta^{*T}] + E[YA_2\delta(\beta^T Z)\delta^T(\beta^T Z)A_2^T]. \end{aligned} \tag{A.2}$$

On the other hand, we have

$$\begin{aligned} E(YZZ^T) &= E\{Y[Z^* + A_2\delta(\beta^T Z)][Z^* + A_2\delta(\beta^T Z)]^T\} \\ &= E(YZ^*Z^{*T}) + E[YZ^*\delta^T(\beta^T Z)A_2^T] + E[YA_2\delta(\beta^T Z)Z^{*T}] + E[YA_2\delta(\beta^T Z)\delta^T(\beta^T Z)A_2^T]. \end{aligned} \tag{A.3}$$

For $E(YZZ^T) = E(YU_\beta U_\beta^T)$ and $E(YZ^*Z^{*T}) = E(YU_\beta^* U_\beta^{*T})$ to imply each other, we see from (A.2) and (A.3) that it remains to show

$$E[YZ^*\delta^T(\beta^T Z)A_2^T] = E[YU_\beta^*\delta^T(\beta^T Z)A_2^T] \tag{A.4}$$

and

$$E[YA_2\delta(\beta^T Z)Z^{*T}] = E[YA_2\delta(\beta^T Z)U_\beta^{*T}]. \tag{A.5}$$

Note that U_β^* is a function of $\beta^T Z$, we have

$$E(Z^*|\beta^T Z, Y) = E(Z^*|\beta^T Z) = E(Z^*|\beta^T Z^*) = U_\beta^* = E(U_\beta^*|\beta^T Z, Y).$$

(A.4) is true because

$$E[Y(Z^* - U_\beta^*)\delta^T(\beta^T Z)A_2^T] = E[YE(Z^* - U_\beta^*|\beta^T Z, Y)\delta^T(\beta^T Z)A_2^T] = 0.$$

(A.5) can be shown similarly.

The second part of Proposition 3, which states that $E(Z^*|\beta^T Z^*)$ is linear in $\beta^T Z^*$, has been proved in [15]. □

References

[1] R.D. Cook, On the interpretation of regression plots, Journal of the American Statistical Association 89 (1994) 177–189.
[2] R.D. Cook, Graphics for regressions with a binary response, Journal of the American Statistical Association 91 (1996) 983–992.
[3] R.D. Cook, Principal Hessian directions revisited (with discussion), Journal of the American Statistical Association 93 (1998) 84–100.
[4] R.D. Cook, Regression Graphics: Ideas for Studying Regressions Through Graphics, first ed., Wiley, New York, 1998.
[5] R.D. Cook, B. Li, Dimension reduction for conditional mean in regression, The Annals of Statistics 30 (2002) 455–474.
[6] R.D. Cook, L. Li, Dimension reduction in regression with exponential family predictors, Journal of Computational and Graphical Statistics 18 (2009) 774–791.
[7] R.D. Cook, S. Weisberg, Discussion of sliced inverse regression for dimension reduction, Journal of the American Statistical Association 86 (1991) 316–342.
[8] R.D. Cook, X. Yin, Dimension reduction and visualization in discriminant analysis (with discussion), Australian & New Zealand Journal of Statistics 43 (2001) 147–199.
[9] Y. Dong, B. Li, Dimension reduction for non-elliptically distributed predictors: second order methods, Biometrika 97 (2010) 279–294.
[10] Y. Dong, L.P. Zhu, A note on sliced inverse regression with missing predictors, Statistical Analysis and Data Mining 5 (2012) 128–138.
[11] M.L. Eaton, A characterization of spherical distributions, Journal of Multivariate Analysis 41 (1986) 1–31.
[12] K.C. Li, Sliced inverse regression for dimension reduction (with discussion), Journal of the American Statistical Association 86 (1991) 316–342.
[13] K.C. Li, On principal Hessian directions for data visualization and dimension reduction: another application of Stein's lemma, Journal of the American Statistical Association 87 (1992) 1025–1039.
[14] B. Li, Y. Dong, Dimension reduction for non-elliptically distributed predictors, The Annals of Statistics 37 (2009) 1272–1298.
[15] B. Li, Y. Dong, An iterative algorithm for dimension reduction with non-elliptically distributed predictors, 2012. Unpublished manuscript.
[16] K.C. Li, N. Duan, Regression analysis under link violation, The Annals of Statistics 17 (1989) 1009–1052.
[17] Y. Li, L.X. Zhu, Asymptotics for sliced average variance estimation, The Annals of Statistics 35 (2007) 41–69.
[18] X. Liu, A. Srivastava, K. Gallivan, Optimal linear representations of images for object recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 26 (2003) 662–666.
[19] J.A. Nelder, R. Mead, A simplex method for function minimization, Computational Journal 7 (1965) 308–313.
[20] Y. Xia, H. Tong, W.K. Li, L.X. Zhu, An adaptive estimation of optimal regression subspace, Journal of the Royal Statistical Society. Series B 64 (2002) 363–410.
[21] Z. Ye, R. Weiss, Using the bootstrap to select one of a new class of dimension reduction methods, Journal of the American Statistical Association 98 (2003) 968–979.
[22] X. Yin, E. Bura, Moment based dimension reduction for multivariate response regression, Journal of Statistical Planning and Inference 136 (2006) 3675–3688.
[23] X. Yin, R.D. Cook, Dimension reduction for the conditional kth moment in regression, Journal of the Royal Statistical Society. Series B 64 (2002) 159–175.
[24] X. Yin, B. Li, R.D. Cook, Successive direction extraction for estimating the central subspace in a multiple-index regression, Journal of Multivariate Analysis 99 (2008) 1733–1757.
[25] L.X. Zhu, K.W. Ng, Asymptotics of sliced inverse regression, Statistica Sinica 5 (1995) 727–736.